MINIREVIEW

A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses

Pier Luigi Buttigieg[1,2,3] & Alban Ramette[1,4]

[1]HGF-MPG Group for Deep Sea Ecology and Technology, Bremerhaven, Germany; [2]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany; [3]MARUM Center for Marine Sciences, Bremen, Germany; and [4]Max Planck Institute for Marine Microbiology, Bremen, Germany

Correspondence: Pier Luigi Buttigieg, HGF-MPG Group for Deep Sea Ecology and Technology, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany. Tel.: +49 421 2028 984; fax: +49 421 2028 690; e-mail: pbuttigi@mpi-bremen.de

Present address: Alban Ramette, Institute of Social and Preventive Medicine (ISPM), University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland

Received 16 June 2014; revised 30 September 2014; accepted 6 October 2014. Final version published online 5 November 2014.

DOI: 10.1111/1574-6941.12437

Editor: Gerard Muyzer

Keywords
multivariate statistics; online resource; interactive guide; complex data.

Abstract

The application of multivariate statistical analyses has become a consistent feature in microbial ecology. However, many microbial ecologists are still in the process of developing a deep understanding of these methods and appreciating their limitations. As a consequence, staying abreast of progress and debate in this arena poses an additional challenge to many microbial ecologists. To address these issues, we present the GUide to STatistical Analysis in Microbial Ecology (GUSTA ME): a dynamic, web-based resource providing accessible descriptions of numerous multivariate techniques relevant to microbial ecologists. A combination of interactive elements allows users to discover and navigate between methods relevant to their needs and examine how they have been used by others in the field. We have designed GUSTA ME to become a community-led and -curated service, which we hope will provide a common reference and forum to discuss and disseminate analytical techniques relevant to the microbial ecology community.

Introduction

Multivariate statistical analyses are typically used to summarise high-dimensional data, test hypotheses involving multiple response variables, and examine relationships between large sets of variables (Legendre & Legendre, 1998; Härdle & Simar, 2007). The use of multivariate analyses is supplanting 'simple' descriptive analyses across ecology (see James & McCulloch, 1990 and Økland, 2007 for comment) and has become common in microbial ecology, where complex, multidimensional data sets abound (e.g. Ramette, 2007; Bertics & Ziebis, 2009; Frossard *et al.*, 2012; Thioulouse *et al.*, 2012; Hartmann *et al.*, 2013; Rivers *et al.*, 2013). Indeed, numerous software tools used by microbial ecologists implement multivariate analysis techniques and have been recommended as standard components of, for example, microbiome analysis (Kuczynski *et al.*, 2012) and environmental studies (e.g. Zinger *et al.*, 2012). Notable examples include the MOTHUR software (Schloss *et al.*, 2009), the Quantitative Insights Into Microbial Ecology (QIIME) platform (Caporaso *et al.*, 2010), the PHYLOSEQ package (McMurdie & Holmes, 2013) and the Biodiversity Virtual e-Laboratory (BIOVEL; http://www.biovel.eu/) project. While such developments may lead one to conclude that standard statistical recipes and 'workflows' now exist for microbial ecology data, it is vital to recognise that gauging the appropriateness of a given technique to the data and phenomena under investigation is not necessarily a 'cut and dried' affair.

Firstly, it is essential to recognise that the application of statistical techniques to ecological data is the focus of a living field of study: numerical ecologists and statisticians routinely re-evaluate the properties and limitations of even well-known techniques in relation to ecological

FEMS MICROBIOLOGY ECOLOGY

needs. For example, Legendre (2005b) recently re-examined the value of the Kendall coefficient of concordance (W) in determining species associations in field survey data. Treating species as the 'judges' native to W's conceptual formulation allows the identification of species groups with similar 'opinions' (gauged by their variable values) which may be used as indicators of a given ecological phenomenon; however, Legendre describes several important caveats to the statistic's use in ecology, as not all variables are suited to its assumptions. Similarly, Warton & Hudson (2004) compared the effectiveness of the well-known multivariate analysis of variance (MANOVA) to approaches that rely on the calculation of dissimilarities between sampling units rather than analyse abundance data directly. These authors present a developed case suggesting that the use of dissimilarity-based approaches should be questioned and that alternatives may bring several advantages in generalisation and extensibility. Aside from re-evaluation, proposals of new techniques and adaptations of existing techniques are steadily encountered. For example, Anderson (2001) developed a nonparametric multivariate analysis of variance approach which is argued to be better-suited to ecological data while Zou *et al.* (2006) proposed a form of principal components analysis suited to the sparse data sets generated by, for example, genomic sequencing technologies. Approaches to meaningfully transform ecological data sets for ordination (Legendre & Gallagher, 2001), new ordination approaches (e.g. Pavoine *et al.*, 2004) and methods to systematically assess the impact of rare phylotypes on analytical results (Gobet *et al.*, 2010) provide other examples of relatively recent developments in ecologically oriented multivariate analysis. As they emerge, new techniques which show promise in an empirical setting often require review from expert statisticians to be fully understood. One example features the work of Borcard & Legendre (2002), who proposed a variant of the well-known principal coordinates analysis to detect and characterise spatial structures in ecological data across all scales. In response to these authors' call for more thorough mathematical appraisal of their technique, Dray *et al.* (2006) developed supporting theory and connected the original method to a broader set of autocorrelation functions. From the above examples, it is clear that users of multivariate statistical techniques in microbial ecology must stay abreast of a steadily developing body of work involving a wide range of expertise.

Secondly, to make informed methodological choices, users must be aware of the key debates that emerge in the multivariate analysis of ecological data. For example, a multi-year discussion concerning the analysis of beta diversity using distance-based and 'raw data' approaches recently unfolded in the journal *Ecology* (Legendre, 2005a; Laliberté, 2008; Legendre *et al.*, 2008; Pélissier *et al.*, 2008; Tuomisto & Ruokolainen, 2008, 2006). Distance and dissimilarity measures, such as the well-known Bray–Curtis dissimilarity or Jaccard index, are conceptually appealing as they can address issues such as the handling of the double zero problem: accounting for the fact that observed absences (or zero abundances) of several ecological entities across the same sampling units are not necessarily indicators of similarity between those entities. However, the use of these measures introduces dependencies between objects (e.g. sites, samples, or experimental units) which may violate key assumptions of regression-type analyses and may not deliver as much power as an examination of 'raw' presence–absence or abundance data. On another front, Warton *et al.* (2012) demonstrated that (dis)similarity-based methods confound the mean–variance relationships characteristic of abundance (or other count-based) data. These authors call for greater emphasis to be given to model-based approaches, citing methods based on generalised estimating equations (Warton, 2011) and an original method named constrained additive ordination (Yee, 2006) as examples. Similar debate also surrounds aspects of experimental and sampling design, such as the issue (or, as some contend, nonissue) of pseudoreplication in ecological investigations (Hurlbert, 1984, 2004, 2009; Oksanen, 2001, 2004; Cottenie & De Meester, 2003; Coss, 2009; Koehnle & Schank, 2009; Schank & Koehnle, 2009; Prosser, 2010). While some insist that replication of treatments (or environmental contexts) across 'truly' independent sampling or experimental units must occur to draw valid conclusions, others argue that this may not be an achievable, or even necessary, goal in ecological investigations. The contemporary and faceted nature of such debates presents another challenge to the effective and duly cautious application of powerful analytical methods in microbial ecology.

Lastly, the harmonisation of canonical ecological theory with microbial ecology is ongoing (e.g. Prosser *et al.*, 2007; Ramette, 2007) and faces the challenge of keeping pace with new molecular techniques, sequencing technologies and ecological sampling strategies both on global (Rusch *et al.*, 2007; Karsenti *et al.*, 2011; Zinger *et al.*, 2011) and on local scales (e.g. Kuczynski *et al.*, 2012; Böer *et al.*, 2009; Zhou *et al.*, 2013). Zinger *et al.* (2012) underscored this issue as well as its connection to the use of new statistical techniques in the field of aquatic microbial ecology.

The popularity of multivariate analyses is continuing to increase and their application to microbial ecological data has become technically simplified; however, a developed and up-to-date understanding of their properties and limitations is still not widespread in the community. As a result, many microbial ecologists who are not equipped with deep numerical training face a 'black box' approach

to multivariate analysis and the associated risks of misapplying techniques or misinterpreting results. Reviewers, too, often face uncertainty in evaluating whether researchers have performed appropriate analyses and produced fair interpretations of their results. To support and promote the constantly developing understanding of multivariate analyses in microbial ecology, we present the GUide to STatistical Analysis in Microbial Ecology (GUSTA ME; http://mb3is.megx.net/gustame) – an online, dynamically updated resource with content tailored to the needs of the microbial ecology community.

GUSTA ME: a living reference for multivariate statistics

Periodic reviews of multivariate statistics targeting the uninitiated life scientist (e.g. Ramette, 2007; Jombart *et al.*, 2009) are helpful primers for microbial ecologists, however, must be limited in depth to achieve sufficient breadth. In contrast, seminal textbooks (e.g. Legendre & Legendre, 1998; Borcard *et al.*, 2011; Legendre & Legendre, 2012) offer great depth and breadth, but emerge with low frequency and are rarely targeted to microbial ecologists, who are often confronted with data sets which require specific statistical treatment. We designed GUSTA ME as a compromise: a 'living', web-based, and community-reviewed resource containing descriptions of both established and novel multivariate techniques, specifically curated for their relevance to microbial ecology. Where appropriate, GUSTA ME also discusses the debates noted above – such as that surrounding the issue of pseudoreplication – at greater length. We believe this resource will assist microbial ecologists in navigating the initially daunting field of multivariate analysis by directing them to techniques relevant to their investigations and interests through interactive interfaces. GUSTA ME comprises a collection of interlinked, high-level summaries of multivariate methods (henceforth, 'end points') which users may access (1) directly, (2) by following a series of questions presented by a 'wizard', (3) by following a 'walkthrough' which reflects the analytical procedures used in an existing study or (4) by browsing GUSTA ME's visualisation library (Fig. 1). Below, these components of GUSTA ME as well as its community-led development model are described.

Reference pages as end points

GUSTA ME's core comprises high-level descriptions of a range of multivariate techniques. As these reference pages are arrived at through user interaction, we refer to them as 'end points'. End points avoid technical and formalised

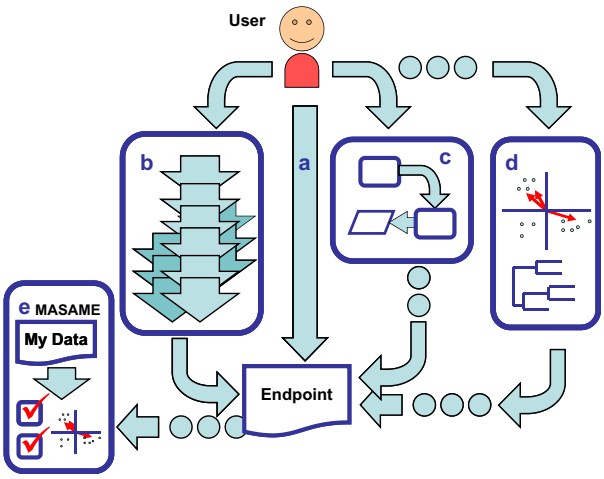

Fig. 1. A user may access or discover GUSTA ME's high-level descriptions of multivariate methods (end points) in a number of ways: (a) should a user know of the method, they may directly navigate to the relevant end point through direct links or via a search function; (b) should the user require guidance, a wizard may be used to guide the user to an end point appropriate to their needs (see text for details); (c) users may browse through walkthroughs (see text) to observe how others have used multivariate methods and navigate to methods that interest them; or (d) users may browse GUSTA ME's visualisation library to select methods based on their visual output. (e) From selected end points, users have the option to launch a MASAME application to perform the featured method (as well as supplementary methods such as data transformations) on their own or preloaded data through an interactive, user-friendly web-page.

mathematical descriptions as far as possible, aiming instead to clarify the conceptual basis of each method, its limitations, and the meaning of its results. Classical techniques, such as canonical correspondence analysis (CCA) and nonmetric multidimensional scaling (NMDS), are included as well as techniques which are relatively new or show significant potential in the field. Examples of the latter category include distance-based redundancy analysis (db-RDA; Legendre & Anderson, 1999) and principal coordinates of neighbour matrices (PCNM; Borcard & Legendre, 2002). Each end point describes the main principles of these methods as well as their key assumptions and output. Commentary on the statistical and ecological interpretation of each endpoint's results is also included as well as warnings emphasising common pitfalls associated with each method. General warnings which refer to common risks (e.g. multiple testing, multicollinearity, data dredging) are explained at greater length in dedicated pages which are linked to end points and intervene, as appropriate, during the course of 'wizards' (see below). Finally, links to references pertinent to each method are provided on their respective description page. Figure 2 illustrates how a user may navigate to the various

components of the guide from a given endpoint, depending on their input and interaction.

Wizards

The interactivity of GUSTA ME is primarily offered through 'wizards': user-interface agents that partition difficult or complex tasks into a linear series of comparatively simple steps. User input determines the succession of these steps and the outcome of the overall task (Dryer, 1997). GUSTA ME's wizards comprise a hierarchical succession of simple questions which approximate the decision-making process of a data analyst. Dependent on their answers, users will be directed to consider a technique or set of techniques which would best match their needs. Dryer (1997) noted that wizards are best suited to tasks whose outcome can be determined by following a predetermined prescription or recipe. Consequently, GUSTA ME's wizards are only able to suggest a single end point when there is a (relatively) clear prescription for an analytical problem. When this is not the case or when the answers required of the user are too technical in nature (i.e. they presuppose knowledge which the target users of the guide are not expected to have), a GUSTA ME wizard will present a brief description of methods that *may* suit the user's needs and will link to their end points. It is then left to the user to familiarise themselves with the techniques suggested and make an informed choice or to interact with other users via GUSTA ME's community forum (see below). Wizards will be adapted as new end points are added to GUSTA ME or in response to community input.

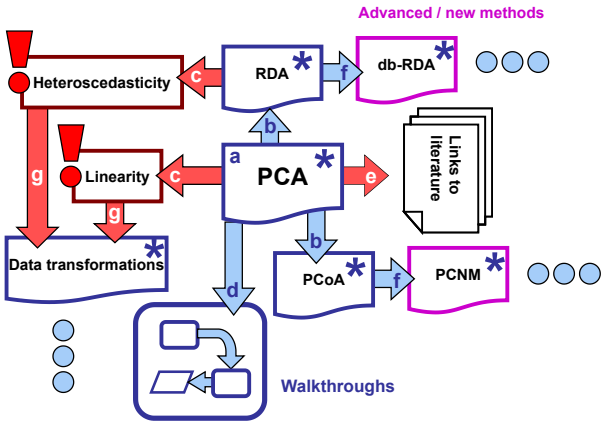

Fig. 2. End points are linked to relevant material across GUSTA ME, allowing users to deepen and broaden their understanding. As an example, the end point for principal components analysis (PCA; a) is shown linked to: (b) end points of related techniques such as those for principal coordinates analysis (PCoA) and redundancy analysis (RDA); (c) pages describing warnings that are associated with the technique; (d) walkthroughs (see text for description) that feature the method described; and (e) links to relevant literature. Further, each page linked to a given end point is also linked to other material such as (f) relatively new or advanced techniques and (g) potential approaches to contend with warnings. Online applications in the MASAME suite (see text) allow users to apply the method to their own data are launchable from selected pages across GUSTA ME (here, indicated by an asterisk).

Example-based learning through 'walkthroughs'

Peer-reviewed studies in microbial ecology which have employed multivariate techniques serve as important exemplars for the community. A section of GUSTA ME is dedicated to the capture of such exemplars and the visualisation of their analytical methodology as approachable, interactive flowcharts dubbed 'walkthroughs'. Key steps or methods included in the walkthrough are linked to the relevant end point(s), connecting users to GUSTA ME's content and curated reference material. This collection of methodological summaries provides an opportunity for microbial ecologists to examine, in an example-based manner, under what circumstances and with what forms of data multivariate analyses have been used by the community. Sections of GUSTA ME's community forums are dedicated to the discussion of these walkthroughs, and users may contribute their own walkthroughs to the guide.

Visualisation libraries

Data visualisation is a major outcome of many multivariate analyses and is instrumental in rendering high-dimensional data into a form that humans can grasp. Graphs, charts and plots native to many multivariate techniques are designed to be readily interpretable by analysts and nonanalysts alike. The recognition of an effective visualisation may occur without deep knowledge of the underlying mathematical basis of a technique. GUSTA ME features a library of visualisations which may be browsed to guide users to a technique or family of techniques which may deliver a useful representation of their data. Visualisations link to an appropriate end point or, when user input is required, to a wizard.

Analysis applications – the MASAME suite

Selected pages across GUSTA ME include links to interactive analysis applications which allow users to perform the technique or procedure discussed on that page, either on their own data sets (which may be uploaded as comma-separated-value files) or on preloaded

example data. Collectively, we refer to these applications as the Multivariate AnalysiS Applications for Microbial Ecology (MASAME) suite. MASAME applications are accessed through user-friendly web-pages, rendered by the *shiny* package (RStudio Inc., 2014), which call upon numerous functions from well-known packages belonging to the statistical programming environment and language, R (R Development Core Team, 2014). For example, (partial) RDA, (partial) CCA, NMDS, and PCNM methods from the *vegan* package (Oksanen *et al.*, 2013) are combined with supporting functions which allow data transformations using standard and ecologically meaningful methods (after Legendre & Gallagher, 2001), plotting, and download functionality on a single webpage. Users need not know the R language, as point-and-click interfaces are common to all MASAME applications. Such tools add a practical complement to GUSTA ME's review of multivariate analysis techniques and are easily enhanced to address new needs as they arise.

Community involvement and development

As described above, statistical methods relevant to microbial ecologists are constantly emerging, either through novel development or through adaptation from other domains. A single working group is likely to overlook or only partly represent developments which may be of great use or importance. Thus, GUSTA ME and MASAME are linked to on-line forums where users may comment on their content, discuss the methods featured, suggest revisions, post critiques, and note alternative views. User input will allow these resources to grow based on the needs of the microbial ecology community and will offer a gateway for new contributors, moderators, and editors with additional analytical expertise to join and enhance the guide. This will be particularly useful in popularising less well-known techniques that are better-suited to specific scenarios in microbial ecology relative to more classical methods. By providing such a service, we hope to foster both a community-curated reference and a forum for microbial ecologists to share their evaluations of multivariate techniques. We hope that as consensuses are reached, alternatives put forth, and gaps in the domain's analytical and theoretical repertoire highlighted, GUSTA ME will serve to encourage analytical consistency and transparency across microbial ecology.

Usage examples

Below, we describe three usage scenarios of GUSTA ME from the perspectives of a doctoral student in search of a method to explore their multivariate data, a principal investigator formulating a project proposal, and a reviewer harbouring concerns about a manuscript's analytical methods.

The student

A doctoral student wishes to explore a priori groupings in a data set containing sampling sites as objects and OTU relative abundances as variables; however, the student is unsure where to begin. Using the 'Explore data' starting point on GUSTA ME's home page, the student enters a wizard and is prompted to consider screening their data for, among other features, the presence of outliers, multicollinearity, and (multivariate) normality (see e.g. Zuur *et al.*, 2010). The student navigates to description pages for these tasks and is able to use applications from the MASAME suite to evaluate and preprocess their data.

With their data screened and appropriately preprocessed, the student then proceeds to the data exploration wizard. The wizard presents a warning page describing the unfortunately all-too-common mistake of 'data dredging' or 'P-hacking' (Nuzzo, 2014). Suitably warned, the student then proceeds and, after determining that the differences between their samples are of prime interest, chooses the path of analysis based on (dis)similarity matrices. A page is presented which lists and briefly describes the aims of several (dis)similarity-based ordination, clustering, and hypothesis testing methods. The student decides to attempt an NMDS ordination complemented with an Analysis of Similarity (ANOSIM; Clarke, 1993) hypothesis test (avoiding data dredging) and navigates to these methods' end points. Somewhat uncertain about the nature of (dis)similarity measures, the student follows a link present on both the end points and the exploration wizard to an end point and another wizard dealing with (dis)similarity measures. Through this excurse, the student is able to make an informed choice regarding the most appropriate measure to use with their (dis)similarity-based method of choice. Returning to the NMDS and ANOSIM end points and using them much like short review articles, the student becomes familiar with the requirements and limitations of the candidate methods.

Satisfied that the candidate methods are appropriate, the student launches the methods' MASAME applications via links on each end point. They upload their data and, following the advice in each end point, interactively adjust the methods' parameters to suitable values. As the results look promising, the student explores some of the relevant literature listed on each methods' end point to deepen their understanding. The student also browses GUSTA ME's community forum to familiarise themselves with frequently asked questions concerning how their

method of choice is best applied to microbial ecology data. Ultimately, the student publishes a study skilfully employing multivariate analyses and contributes a walkthrough based on their study to GUSTA ME.

The principal investigator

A principal investigator, who has an introductory familiarity with multivariate statistical methods, is designing an investigation to assess whether energy availability drives microbial community change in a little-studied environment or whether changes are simply a function of spatial distance. Curious as to which multivariate methods others have used to approach similar questions, the PI browses GUSTA ME's walkthroughs and interacts with their components to quickly learn more. Encountering walkthroughs based on Bienhold *et al.* (2012) and Kopp *et al.* (2012) which feature the use of RDA, variation partitioning (Legendre, 2005a, 2007; Peres-Neto *et al.*, 2006), path analysis (Wright, 1934), and PCNM, the PI familiarises themselves with the interplay of the central and ancillary methods involved in these studies and, supported by GUSTA ME's end points and warnings, augments their initial project proposal. In particular, the PI realises the central importance of aligning their sampling design and replication strategy to their proposed analytical approaches in order to arrive at valid conclusions. The PI also feels that several important methods are not noted in GUSTA ME, and uses GUSTA ME's community forum to post requests for enhancement and adds contact details of relevant experts. The GUSTA ME editors contact these experts and invite them to create and manage end points and wizards aligned with their expertise.

The reviewer

A reviewer is unsure about the appropriateness of a parametric multivariate hypothesis test, multivariate analysis of variance (MANOVA), in a manuscript. The reviewer uses GUSTA ME's search function to locate the relevant end point and ensures that the manuscript adequately reports if the method's key assumptions have been met. The reviewer finds that the authors have not reported if their data have been appropriately transformed to meet the assumptions of near-normality, linearity, and homogeneity of covariances. Further, the authors have not reported if they have screened their data for outliers. Fortunately, the study's authors have made their data available for review. The reviewer downloads the data, uploads it to the MASAME data screening applications, and concludes that it is very unlikely that this parametric hypothesis test can be applied in its basic form. Further, while reading the MANOVA end point, the reviewer is directed to an end point describing permutational, nonparametric MANOVA (NPMANOVA or PERMANOVA; Anderson, 2001). The reviewer suggests that the authors explore this method or rigorously justify their use of MANOVA.

Conclusion & outlook

GUSTA ME is an interactive 'living' review of multivariate analyses with specific relevance to the microbial ecology community. Its content offers an accessible resource for teaching and reference, while its implementation allows users to quickly locate and focus their efforts on analytical approaches pertinent to their investigations. We recognise that the current state of GUSTA ME is but a starting point for a more comprehensive solution; however, we are confident that, even in its current form, it will provide a useful resource for microbial ecologists wishing to delve deeper into multivariate statistics. As it further develops, GUSTA ME has the potential to become a focal repository for accessible analytical knowledge and debate in microbial ecology, wherein methods that have become central to ecology, as well as their criticisms (e.g. Warton *et al.*, 2012), may be easily explored. In the near future, GUSTA ME and MASAME will be integrated into the MicroB3 Information System (MicroB3 IS; www.microb3.eu) for further development. The MicroB3 IS, based on the megx.net web platform (Kottmann *et al.*, 2010), will serve as a multicomponent information system for European marine microbial genomics, integrating genomic and environmental data with an array of tools and services for the global research community. GUSTA ME and MASAME will complement the system's data management, integration, and processing modules by providing support in the analysis of integrated data; however, both resources may also be used independently. Guided by user input, we will implement the user-feedback mechanisms and editorial policies required to allow community-led development of this resource. We hope these efforts will promote the usage, discussion, and development of multivariate statistical approaches in microbial ecology and look forward to the involvement of the community in this endeavour.

Acknowledgements

This work is a component of the Micro B3 project and is funded by the European Union's Seventh Framework Programme (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under grant agreement no 287589. AR is funded by the Max Planck Society. The authors declare no conflict of interests.

References

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**: 32–46.

Bertics VJ & Ziebis W (2009) Biodiversity of benthic microbial communities in bioturbated coastal sediments is controlled by geochemical microniches. *ISME J* **3**: 1269–1285.

Bienhold C, Boetius A & Ramette A (2012) The energy-diversity relationship of complex bacterial communities in Arctic deep-sea sediments. *ISME J* **6**: 724–732.

Böer SI, Hedtkamp SIC, van Beusekom JEE, Fuhrman JA, Boetius A & Ramette A (2009) Time- and sediment depth-related variations in bacterial diversity and community structure in subtidal sands. *ISME J* **3**: 780–791.

Borcard D & Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol Model* **153**: 51–68.

Borcard D, Legendre P & Gillet F (2011) *Numerical Ecology with R* (Gentleman R, Hornik K & Parmigiani GG, eds), Springer, New York.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Austral Ecol* **18**: 117–143.

Coss RG (2009) Pseudoreplication conventions are testable hypotheses. *J Comp Psychol* **123**: 444–446.

Cottenie K & De Meester L (2003) Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). *Oikos* **100**: 394–396.

Dray S, Legendre P & Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol Model* **196**: 483–493.

Dryer DC (1997) Wizards, guides, and beyond. *Proceedings of the 2nd International Conference on Intelligent User Interfaces – IUI '97* (Moore J, Edmonds E & Puerta A, eds), pp. 265–268. ACM Press, New York.

Frossard A, Gerull L, Mutz M & Gessner MO (2012) Disconnect of microbial structure and function: enzyme activities and bacterial communities in nascent stream corridors. *ISME J* **6**: 680–691.

Gobet A, Quince C & Ramette A (2010) Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res* **38**: e155.

Härdle W & Simar L (2007) *Applied Multivariate Statistical Analysis*, 2nd edn. Springer, Berlin, Heidelberg.

Hartmann M, Niklaus PA, Zimmermann S, Schmutz S, Kremer J, Abarenkov K, Lüscher P, Widmer F & Frey B (2013) Resistance and resilience of the forest soil microbiome to logging-associated compaction. *ISME J* **8**: 226–244.

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* **54**: 187–211.

Hurlbert SH (2004) On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. *Oikos* **104**: 591–597.

Hurlbert SH (2009) The ancient black art and transdisciplinary extent of pseudoreplication. *J Comp Psychol* **123**: 434–443.

James F & McCulloch C (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annu Rev Ecol Syst* **21**: 129–166.

Jombart T, Pontier D & Dufour A-B (2009) Genetic markers in the playground of multivariate analysis. *Heredity* **102**: 330–341.

Karsenti E, Acinas SG, Bork P *et al.* (2011) A holistic approach to marine eco-systems biology. *PLoS Biol* **9**: e1001177.

Koehnle TJ & Schank JC (2009) An ancient black art. *J Comp Psychol* **123**: 452–458.

Kopp D, Bouchon-Navaro Y, Louis M, Legendre P & Bouchon C (2012) Spatial and temporal variation in a Caribbean herbivorous fish assemblage. *J Coast Res* **278**: 63–72.

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, Waldmann J & Glöckner FO (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38**: D391–D395.

Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D & Knight R (2012) Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* **13**: 47–58.

Laliberté E (2008) Analyzing or explaining beta diversity? Comment. *Ecology* **89**: 3232–3237.

Legendre P (2005a) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol Monogr* **75**: 435–450.

Legendre P (2005b) Species associations: the Kendall coefficient of concordance revisited. *J Agric Biol Environ Stat* **10**: 226–245.

Legendre P (2007) Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *J Plant Ecol* **1**: 3–8.

Legendre P & Anderson MJ (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol Monogr* **69**: 1–24.

Legendre P & Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**: 271–280.

Legendre P & Legendre L (1998) *Numerical Ecology*, 2nd edn. Elsevier, Amsterdam.

Legendre P & Legendre L (2012) *Numerical Ecology*, 3rd edn. Elsevier, Amsterdam.

Legendre P, Borcard D & Peres-Neto P (2008) Analyzing or explaining beta diversity? Comment. *Ecology* **89**: 3238–3244.

McMurdie PJ & Holmes S (2013) phyloseq : an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**: e61217.

Nuzzo R (2014) Scientific method: statistical errors. *Nature* **506**: 150–152.

Økland RHR (2007) Wise use of statistical tools in ecological field studies. *Folia Geobot* **42**: 123–140.
Oksanen L (2001) Logic of experiments in ecology: is pseudoreplication a pseudoissue? *Oikos* **94**: 27–38.
Oksanen L (2004) The devil lies in details: reply to Stuart Hurlbert. *Oikos* **104**: 598–605.
Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH & Wagner H (2013) *vegan: community Ecology Package*. R package version 2.0-7. http://CRAN.R-project.org/package=vegan.
Pavoine S, Dufour A-B & Chessel D (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J Theor Biol* **228**: 523–537.
Pélissier R, Couteron P & Dray S (2008) Analyzing or explaining beta diversity? *Ecology* **89**: 3227–3232.
Peres-Neto P, Legendre P, Dray S & Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* **87**: 2614–2625.
Prosser JI (2010) Replicate or lie. *Environ Microbiol* **12**: 1806–1810.
Prosser JI, Bohannan B & Curtis T (2007) The role of ecological theory in microbial ecology. *Nat Rev Microbiol* **5**: 384–392.
R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/
Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* **62**: 142–160.
Rivers AR, Sharma S, Tringe SG, Martin J, Joye SB & Moran MA (2013) Transcriptional response of bathypelagic marine bacterioplankton to the Deepwater Horizon oil spill. *ISME J* **7**: 2315–2329.
RStudio Inc. (2014) *shiny: Web Application Framework for R*. R package version 0.10.2.1. http://CRAN.R-project.org/package=shiny.
Rusch DB, Halpern AL, Sutton G *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.
Schank JC & Koehnle TJ (2009) Pseudoreplication is a pseudoproblem. *J Comp Psychol* **123**: 421–433.
Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
Thioulouse J, Prin Y & Duponnois R (2012) Multivariate analyses in soil microbial ecology: a new paradigm. *Environ Ecol Stat* **19**: 499–520.
Tuomisto H & Ruokolainen K (2006) Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. *Ecology* **87**: 2697–2708.
Tuomisto H & Ruokolainen K (2008) Analyzing or explaining beta diversity? Reply. *Ecology* **89**: 3244–3256.
Warton DI (2011) Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* **67**: 116–123.
Warton DI & Hudson HM (2004) A MANOVA statistic is just as powerful as distance-based statistics for multivariate abundances. *Ecology* **85**: 858–874.
Warton DI, Wright ST & Wang Y (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol Evol* **3**: 89–101.
Wright S (1934) The method of path coefficients. *Ann Math Stat* **5**: 161–215.
Yee TW (2006) Constrained additive ordination. *Ecology* **87**: 203–213.
Zhou J, Jiang Y-H, Deng Y, Shi Z, Zhou BY, Xue K, Wu L, He Z & Yang Y (2013) Random sampling process leads to overestimation of β-diversity of microbial communities. *mBio* **4**: e00324-13.
Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, Martiny JBH, Sogin M, Boetius A & Ramette A (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6**: e24570.
Zinger L, Gobet A & Pommier T (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21**: 1878–1896.
Zou H, Hastie T & Tibshirani R (2006) Sparse principal component analysis. *J Comput Graph Stat* **15**: 265–286.
Zuur AF, Ieno EN & Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. *Methods Ecol Evol* **1**: 3–14.